DOCUMENT RESUME

ED 373 100 TM 021 995

AUTHOR Robinson, Glen E.; Brandon, David P.
TITLE NAEP Test Scores: Should They Be Used To Compare and Rank State Educational Quality? ERS Concerns in Education.
INSTITUTION Educational Research Service, Arlington, Va.
PUB DATE 94
NOTE 29p.
AVAILABLE FROM Educational Research Service, 2000 Clarendon Boulevard, Arlington, VA 22201 (staff members in subscribing school districts and agencies, $6; nonsubscribers, $12 plus the larger of $3.50 or 10% of purchase price for postage).
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Community Influence; *Comparative Analysis; Demography; *Educational Quality; Family Structure; Grade 8; Junior High Schools; Junior High School Students; National Competency Tests; Parent Education; Poverty; Program Evaluation; *Scores; State Programs; *Testing Programs; Test Results; *Test Use
IDENTIFIERS *National Assessment of Educational Progress; Ranking; *Trial State Assessment (NAEP); Variance (Statistical)

ABSTRACT

Whether state average test scores on the 1992 National Assessment of Educational Progress (NAEP) Trial State Assessment in mathematics should be used to compare and rank states according to the quality or proficiency of their educational programs is examined. Data from eighth graders from the Trial State Assessment are analyzed. Analysis shows that 89 percent of the variation in state average test scores on the NAEP-92 Trial State Assessment can be explained by the combination of four demographic factors over which states have no control: number of parents living at home, parental education, community type, and state poverty rates. Therefore, Trial State Assessment mathematics scores are not valid measures of the relative quality of the states' educational programs but seem rather to reflect the difficulty of the educational tasks confronting the various states. Proposals to adjust ranking by statistically adjusting the scores to reflect state variations in demographic characteristics are meeting with much controversy. Comparisons and rankings on the basis of these scores do not provide fair measures of educational quality. Seven figures, and four more in an appendix, present analysis results. (Contains 17 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 373 100

CONCERNS IN EDUCATION

Presenting and Analyzing Important Educational Issues

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
DEBORAH PERKINS-GOUGH
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

NAEP Test Scores:

Should They Be Used to Compare and Rank State Educational Quality?

by Glen E. Robinson and David P. Brandon

Educational Research Service

BEST COPY AVAILABLE

TM021995

Educational Research Service
2000 Clarendon Boulevard
Arlington, Virginia 22201
Phone: (703) 243-2100
Fax: (703) 243-8316

The Information Source for School Decisions

Sponsored by:

- American Association of School Administrators
- American Association of School Personnel Administrators
- Association of School Business Officials
- Council of Chief State School Officers
- National Association of Elementary School Principals
- National Association of Secondary School Principals
- National School Public Relations Association

Educational Research Service is an independent, nonprofit corporation serving the research and information needs of the nation's school systems, their administrative staffs, teachers, school boards, and the public.

This publication, *NAEP Test Scores: Should They Be Used to Compare and Rank State Educational Quality?*, is one of the *ERS Concerns in Education* series which is part of the continuing effort of ERS to provide reliable data, information, analyses, and commentary on important problems and issues of concern to persons and organizations responsible for making decisions related to American education.

The observations, analyses, and interpretations contained in the *ERS Concerns in Education* series are those of the authors and do not necessarily represent those of the Educational Research Service or any of its sponsoring organizations.

About the Authors: Glen E. Robinson is President and Director of Research of the Educational Research Service. David P. Brandon is an ERS Research Specialist.

ORDERING INFORMATION: A copy of this *ERS Concerns in Education* has been mailed to each ERS subscribing school district and agency as part of its annual order for ERS services. Staff members in subscribing school districts and agencies may purchase additional copies at the 50% discount price of $6.00. Nonsubscribers may purchase copies at the base price of $12.00.

All mail orders must be accompanied by check or purchase order. Phone orders accepted with Visa or MasterCard. Please add the larger of $3.50 or 10% of purchase price for postage and handling.

Stock #0182

Copyright © 1994 Educational Research Service

TM

NAEP Test Scores:
Should They Be Used to Compare and Rank State Educational Quality?

by Glen E. Robinson and David P. Brandon

Contents

Overview ..... *ii*
Introduction 1
Purpose of this Analysis 2
Data Sources and Focus of the Analysis 3
Student Demographic Variables Used in the Analysis 3
Student Demographic Variables Associated with Lower Scores on the 1992 Math NAEP 4
Combined Effects Explain 89 Percent of State Variation 6
Observations About Rank Order Data 11
Variations in State Participation 11
Major Effects of Non-Response Bias 12
Other Supporting Analyses 12
Relationship of Demographic Variables to District-by-District Comparisons 13
State NAEP Score Adjustment Controversy 13
Measuring the Difficulty of the Educational Task 15
State Data Used in this Analysis 15
Summary and Conclusions 15
Technical Note 18
References 19
Appendix 21

Overview

This *Concerns in Education* examines the important question of whether state average test scores on the 1992 National Assessment of Educational Progress (NAEP) Trial State Assessment in mathematics should be used to *compare* and *rank* the states according to the quality or proficiency of their educational programs. Other uses or purposes of NAEP state or national assessments are not addressed in the analysis.

Findings:

- This analysis shows that 89 percent of the variation in state average test scores on the NAEP-92 Trial State Assessment in mathematics can be explained by the combined effects of four *demographic* factors over which schools have no control — number of parents living at home, parent(s)' education, community type, and state poverty rates.

- Thus, NAEP-92 Trial State Assessment math scores are shown *not* to be valid measures of the relative quality or proficiency of mathematics instruction among the states and, therefore, should *not* be used to *compare* or *rank* states according to the quality or proficiency of their educational programs.

- Rather than accurately measuring differences in the quality or proficiency of the states' educational programs, the NAEP Trial State Assessment test scores appear to more accurately reflect differences in the *difficulty of the educational tasks* confronting the various states.

- Other analysts have found similar strong relationships between demographic variables and student average scores, on statewide tests, among *local school districts* within states.

- In addition to the strong demographic influences on NAEP-92 Trial State Assessment math scores, certain *non-demographic* variables may have substantial influence on NAEP state test score averages and thus seriously affect state *rankings* and state *comparisons*.

- Proposals to provide "fairness" in state NAEP test score rankings by statistically adjusting the scores to reflect state variations in factors such as school resources, ethnic groups, and specific demographic characteristics of the student population have met with much controversy.

- The findings of this analysis regarding the strong influence of demographic factors on state NAEP assessment scores indicate the need to focus attention, efforts, and resources toward helping states address the *difficulty of the educational tasks* they face, rather than using NAEP Trial State Assessment data to inappropriately and unfairly *compare* and *rank* the states on presumed differences in the *quality* of their educational programs.

Introduction

In the late 1980s, the U.S. Congress amid much controversy expanded the National Assessment of Educational Progress (NAEP) to provide state-by-state test score data on a trial basis for states volunteering to participate. This analysis examines the 1992 NAEP Trial State Assessment data to determine whether or not state average NAEP test scores in mathematics should be used to *compare* and *rank* the states on the relative quality, or proficiency, of their educational programs.

The original National Assessment of Educational Progress was initiated in the late 1960s to measure changes in student achievement in the nation, in much the same way that the Consumer Price Index measures changes in the cost of living in the nation's economy.

Such NAEP assessments were intended to provide the nation with descriptors showing what proportion of 9, 13, or 17 year olds in the entire population of the U.S. were able to answer specific kinds of academic questions or perform specific academic tasks. Thus, these indicators were designed to show improvements or declines in student learning in the nation as a whole and for specific demographic groups (Haertel 1991, 17).

From the inception of the NAEP, great pains were taken to design NAEP assessments so that it would be impossible to compare state scores or even to obtain average scores for a specific state. Heated controversy and debates focused on how to ensure that the NAEP would never be used to compare one state with another state. Edward H. Haertel of Stanford University states that:

> Using NAEP for accountability purposes, though, to compare one state with another, for example, or to judge performance against some "standard" or criterion level was carefully avoided (1991, 17).

But the educational reform movement of the 1980s brought calls for greater accountability, and created great pressure to make NAEP assessments perform *accountability* measurement functions, including state-by-state rankings and comparisons of student academic performance.

Responding to such pressure, the Congress in 1988 authorized NAEP voluntary state assessment *trials*. The NAEP Trial State Assessment Program was begun in 1990 with grade 8 tests in mathematics, and with 37 states, the District of Columbia, and two territories participating (Phillips 1991). The trial state assessments were expanded in 1992 to include grade 4 in reading and grades 4 and 8 in mathematics, with 42 states (including D.C.) and two territories participating (Mullis et al. 1993, 5).

As has occurred with other state-level aggregated student test scores such as the SAT, state educational systems have been rated, ranked, and compared based on their state average NAEP scores. In the past, the public media

have often been accused of misusing data such as SAT scores to rank the states according to their presumed relative educational quality (Robinson and Brandon 1992, 2-6). But the public media cannot be blamed for initiating use of state comparisons in the case of NAEP-92 Trial State data.

The NAEP-92 Mathematics Trial State Assessment reports, prepared by the Educational Testing Service under contract to the U.S. Department of Education, provide charts to assist readers in making state comparisons and obtaining state rankings. The two figures in the Appendix (see pages 22 and 23) — reproduced from the *NAEP 1992 Mathematics Report Card for the Nation and the States* — are examples of the ways in which the report encourages and helps readers to make state *comparisons* and obtain *rankings*, which are identified as indicating "significant differences between states in average mathematics proficiency" (Mullis et al. 1993, 14-15).

Purpose of this Analysis

This analysis examines data from the 1992 NAEP Trial State Assessment in mathematics to address the question:

> Do the NAEP Trial State Assessment data provide meaningful and valid comparisons of the relative educational quality, or proficiency, of specific states?

Writing in 1991 concerning the initial NAEP-90 Trial State Assessment, Daniel Koretz, then Senior Social Scientist at the RAND corporation, stated:

> To infer that a difference between two states on the NAEP reflects specific policies or practices, one needs to be able to reject with reasonable confidence other plausible explanations, such as economic or demographic differences (1991, 20).

This analysis will examine certain student demographic variables associated with state-level differences in mathematics as reported in the 1992 NAEP Trial State Assessment to see whether these demographic variables provide plausible explanations for state differences in student NAEP test score averages. It is important to do this before assuming that the variations in NAEP-92 Trial State math scores reflect *meaningful* differences in state educational policies, programs, practices, or proficiencies.

The student demographic variables (that is, population characteristics) associated with *lower* state average NAEP-92 scores in mathematics will be examined to see whether these variables explain a substantial amount of the differences in state average NAEP mathematics scores. Such differences are assumed in the NAEP-92 Trial State Assessment reports to reflect *significant* differences in the relative quality or proficiency of the states' educational programs.

It should be noted that this analysis does not address any of the other uses or purposes of NAEP state or national assessments.

Data Sources and Focus of the Analysis

State NAEP scores and data on student characteristics used in the analysis are taken from the *Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States* produced by the U.S. Department of Education, National Center for Education Statistics, and from the reports of the U.S. Bureau of the Census.

While both 4th and 8th graders participated in the 1992 NAEP Trial State Assessment, only data on 8th graders are used in this analysis because their self-reported answers to demographic questions are expected to be more accurate than those of 4th graders.

The 1992 NAEP includes information on several student-related factors. However, this analysis will deal with only four demographic factors — three factors as reported by NAEP-92 8th-grade test takers in each of the participating states, plus one demographic factor as reported by the U.S. Bureau of the Census for each of the states.

The reasons for concentrating on these four demographic factors are: 1) research on "at-risk" students typically emphasizes demographic factors; 2) educational policy makers and concerned citizens are more likely to attribute objectivity to demographic data; and 3) other research on the NAEP and other large-scale assessments have found links between demographic variables and student test scores. (See Cooley 1993; Drazen 1992; Lapointe et al. 1992; Lazer 1992; Wolf 1992; Pallas et al. 1989.)

The data used in the analysis for each state represent the proportion of the tested students in the state having a particular characteristic. For example, 19 percent of 8th-grade students participating in NAEP-92 in California reported living in a disadvantaged urban community (see Figure 7, page 16).

Student Demographic Variables Used in the Analysis

The student demographic variables (that is, population characteristics or factors) reported by 8th-grade NAEP test takers and used in the analysis include indicators of:

1. number of parents living at home
2. parent(s)' educational background
3. community type.

Figure 1 lists these three demographic variables, and shows the different categories or levels by which each variable is reported in the NAEP-92 data tables (NCES 1993, pp. 73, 83, 702). An asterisk indicates categories or levels of each variable that were negatively correlated with student math scores and were used in this analysis, as shown later in Figure 2 on page 5.

The fourth demographic variable used in the analysis and shown in the figure is the 1992 state *poverty rate* for children ages 5-17 in each state provided by the U. S. Bureau of the Census (1993).

Figure 1.—Demographic Variables Used in this Analysis, Showing Different Categories Reported in NAEP-92 Mathematics Assessment

1. Number of Parents Living at Home: Both parents at home One parent at home* Neither parent at home*	**3. Community Type:** Advantaged urban Disadvantaged urban* Extreme rural Other community type
2. Parent(s)' Educational Background: Don't know Not high school graduate* High school graduate only* Some education after high school College graduate	**4. State Poverty Rate, Ages 5-17:** These are 1992 state poverty rates as reported by the U.S. Bureau of the Census.

* Categories or levels of variable used in this analysis.

SOURCE: National Center for Education Statistics 1993; U.S. Bureau of the Census 1993.

Student Demographic Variables Associated with Lower Scores on the 1992 Math NAEP

Each level or category of the demographic variables shown in Figure 1 was correlated with the state mean scores from the 1992 NAEP mathematics test. Those demographic variable categories having a *negative* correlation with NAEP state-level math scores were used in this analysis and are shown in Figure 2 on page 5. Categories of the variables having *positive* correlations with student math scores, such as "parent(s) college graduate," are not shown in Figure 2 and were *not* used in this analysis.

A negative correlation means that as one variable *increases*, the other *decreases*. For example, as the percentage of children in a state living with only one parent increases, mean state test scores decrease. In other words,

the greater the proportion of children living with a single parent in the state, the lower the mean state NAEP score.

Figure 2 also shows the statistical values for "R" and "R^2" for each of the variable levels listed. The "R" indicates the *direction* (either positive or negative) of the relationship in the correlation, which in these cases is negative for all of the variables listed — for example, -.84 for students living in a home with *only one parent*. The "R^2" (which is the square of the coefficient of correlation "R") is termed the *variance* and describes the *percentage* of variability among the states' NAEP-92 mathematics scores that can be predicted, or accounted for, using each of the variables shown.

For example, 71 percent of the variation in the 1992 state NAEP mathematics test scores can be predicted using only data on percentage of test takers in a state living in a home with *only one parent* (-.84 x -.84 = .7056 = 71%). Likewise, 56 percent of the state differences in NAEP-92 mathematics test scores can be predicted by using the single variable of *state poverty rates* for the population ages 5-17 years.

Figure 2.—Demographic Variable Categories Associated Negatively with State-Level Performance[1] on the 1992 NAEP Mathematics Assessment: Correlation (R) and Percentage of Variability Explained by Each Variable (R^2)

Variable Categories or Levels	Correlation (R)	Percentage of Variability Explained by Variable (R^2)
Neither parent living at home	-.86	74%
One parent living at home	-.84	71%
State 1992 poverty rate, ages 5-17	-.75	56%
Parent(s) not high school graduate	-.68	46%
Living in a disadvantaged urban community	-.55	30%
Parent(s) high school graduate only	-.46	21%
Total Variability Explained by *All* Demographic Variables in Combination[2]		89%

[1]Includes variables with statistically significant negative correlations at the .01 level of testing. Calculated using StatView SE+Graphics software.

[2]This is a combined variable effect, not the sum of the variables listed.

Combined Effects Explain 89 Percent of State Variation

When the effects of the demographic variable categories shown in Figure 2 are *combined* so as to account for any overlapping effects, 89 percent of the differences in NAEP-92 mathematics average scores for the 42 participating states (includes D.C.) can be predicted by the *combined* effects of state variations in the four demographic variables. The 89 percent *combined* effect shown in Figure 2 was calculated by using a multiple regression equation in which the demographic variables were used as "x's" to predict the "y" of NAEP mathematics score averages in 1992. (See Technical Note on page 18.)

Please note that this is *not* a summative procedure. If it were, the total variability would add to more than 100 percent. Rather, the equation used in calculating the *combined* effect takes into account where the variables do and do not overlap each other in predicting the test score for each state.

Figure 3 on page 7 shows the national average student NAEP math score for each of the variable categories used in this analysis. Variables with negative correlations are associated with NAEP mathematics test scores that are lower than the national average student proficiency score of 266.

As described above, the categories of the four demographic factors included in Figure 2, in *combination*, explain 89 percent of the variation in state NAEP-92 math test scores. To illustrate the extent of the influence of the four demographic variables on state mean NAEP math scores in 1992, Figure 4 on page 8 plots two lines. One is the *actual* NAEP scores and the other is the *predicted* NAEP scores using the combined effects of the demographic variables shown in Figure 2.

Observe how closely the *predicted* scores match the *actual* scores. This demonstrates how closely each state's four combined demographic factors predict the state's NAEP mathematics scores for 1992.

Figure 5 on page 9 presents in tabular form the same *actual* and *predicted* state average scores shown in graphic form in Figure 4. The states are listed alphabetically for ease in comparing the *actual* and *predicted* scores and the *actual* and *predicted* rankings for each state. The data indicate that in only 12 of the 42 states did the *predicted* score vary more than 3 points from the state's *actual* NAEP-92 mathematics average test score.

Figure 6 on page 10 shows the same data contained in Figure 5, but with the states listed according to the *rank order* of their average scores in mathematics on the NAEP-92 Trial State Assessment.

Figure 3.— State Average Student Scores on NAEP 1992 Mathematics Assessment, Percent of Students in Demographic Categories, and Correlation with State Average Scores for Categories of Demographic Variables

Variable Variable Category or Level	Student Average NAEP Scores	Percent of Students	Correlation[1] with State Averages
1. Number of Parents Living at Home			
Both parents	274	76%	**.86**[1]
One parent*	260	21%	**-.84**
Neither parent*	246	3%	**-.86**
2. Parent(s)' Educational Background			
Don't know[2]	251	9%	**-.40**
Did not finish high school*	248	8%	**-.68**
High school graduate only*	257	24%	**-.46**
Some education after high school	270	18%	.38
College graduate	280	42%	**.71**
3. Community Type			
Advantaged urban	283	10%	.14
Disadvantaged urban*	238	9%	**-.55**
Extreme rural	267	9%	**.53**
Other	268	72%	.02
4. State Poverty Rate (Ages 5-17)*	—	20%	**-.75**
National 42-State Average in 1992	266	—	—

[1] **Bold** = Statistically significant correlation at the .01 level of significance.

[2] Dropped in this analysis as a non-responsive variable.

*Indicates category or level used in this analysis.

SOURCE: National Center for Education Statistics 1993, pp. 37, 71, 81, 700.

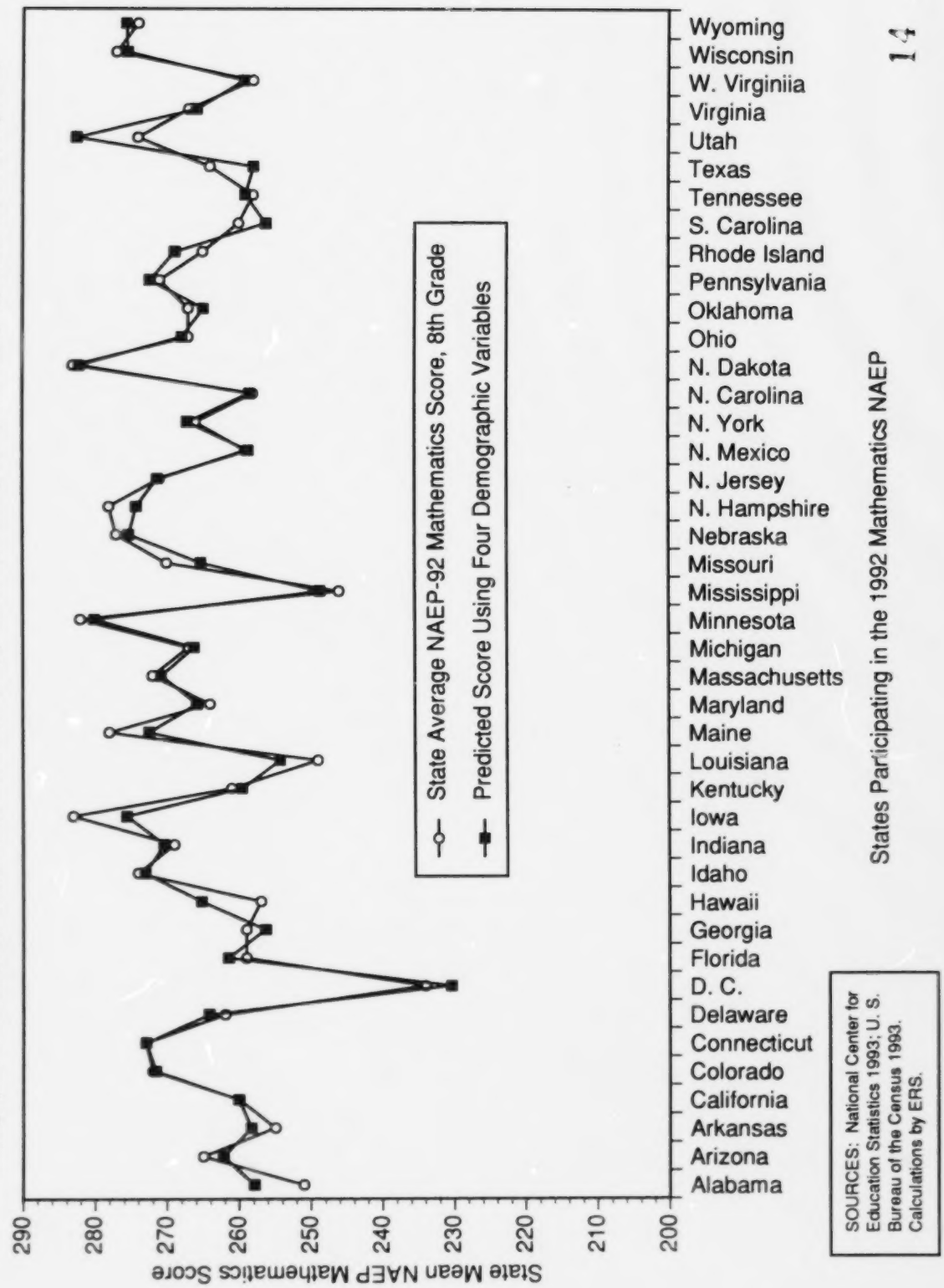


Figure 4. - *Actual* NAEP 1992 State Average Mathematics Scores, 8th Grade, and *Predicted* NAEP Scores Using Combined Effects of Four Demographic Variables (Number of Parents at Home, Parent(s)' Educational Background, Community Type, and State Poverty Rate)

SOURCES: National Center for Education Statistics 1993; U. S. Bureau of the Census 1993. Calculations by ERS.

Figure 5.— Alphabetical Listing of Actual NAEP-92 Mathematics Scores and State Rankings, Plus State Predicted Scores Based on Combined Effects of Four Demographic Variables

Actual 1992			Predicted 1992		
State	Score	Ranking	State	Score	Ranking
Alabama	251	39	Alabama	258	34
Arizona	265	23	Arizona	262	27
Arkansas	255	38	Arkansas	258	34
California	260	29	California	260	29
Colorado	272	12	Colorado	272	11
Connecticut	273	11	Connecticut	273	9
Delaware	262	27	Delaware	264	26
Dist of Colum	234	42	Dist of Colum	230	42
Florida	259	31	Florida	261	28
Georgia	259	31	Georgia	256	38
Hawaii	257	37	Hawaii	265	23
Idaho	274	8	Idaho	273	9
Indiana	269	17	Indiana	270	16
Iowa	283	1	Iowa	276	4
Kentucky	261	28	Kentucky	260	29
Louisiana	249	40	Louisiana	254	40
Maine	278	4	Maine	272	11
Maryland	264	25	Maryland	266	20
Massachusetts	272	12	Massachusetts	271	14
Michigan	267	18	Michigan	266	20
Minnesota	282	3	Minnesota	280	3
Mississippi	246	41	Mississippi	249	41
Missouri	270	16	Missouri	265	23
Nebraska	277	6	Nebraska	275	6
New Hampshire	278	4	New Hampshire	274	8
New Jersey	271	14	New Jersey	271	14
New Mexico	259	31	New Mexico	259	31
New York	266	22	New York	267	19
North Carolina	258	34	North Carolina	258	34
North Dakota	283	1	North Dakota	282	2
Ohio	267	18	Ohio	268	18
Oklahoma	267	18	Oklahoma	265	23
Pennsylvania	271	14	Pennsylvania	272	11
Rhode Island	265	23	Rhode Island	269	17
South Carolina	260	29	South Carolina	256	38
Tennessee	258	34	Tennessee	259	31
Texas	264	25	Texas	258	34
Utah	274	8	Utah	283	1
Virginia	267	18	Virginia	266	20
West Virginia	258	34	West Virginia	259	31
Wisconsin	277	6	Wisconsin	275	6
Wyoming	274	8	Wyoming	276	4

SOURCE: National Center for Education Statistics 1993, p. 37. U. S. Bureau of Census 1993. Calculations by ERS.

Figure 6.— Rank Order Listing of Actual NAEP-92 Mathematics Scores and Predicted Scores Based on Combined Effects of Four Demographic Variables

Actual 1992			Predicted 1992		
State	Score	Ranking	State	Score	Ranking
Iowa	283	1	Utah	283	1
North Dakota	283	1	North Dakota	282	2
Minnesota	282	3	Minnesota	280	3
New Hampshire	278	4	Wyoming	276	4
Maine	278	4	Iowa	276	4
Wisconsin	277	6	Wisconsin	275	6
Nebraska	277	6	Nebraska	275	6
Utah	274	8	New Hampshire	274	8
Wyoming	274	8	Idaho	273	9
Idaho	274	8	Connecticut	273	9
Connecticut	273	11	Maine	272	11
Massachusetts	272	12	Pennsylvania	272	11
Colorado	272	12	Colorado	272	11
New Jersey	271	14	New Jersey	271	14
Pennsylvania	271	14	Massachusetts	271	14
Missouri	270	16	Indiana	270	16
Indiana	269	17	Rhode Island	269	17
Ohio	267	18	Ohio	268	18
Michigan	267	18	New York	267	19
Oklahoma	267	18	Michigan	266	20
Virginia	267	18	Virginia	266	20
New York	266	22	Maryland	266	20
Arizona	265	23	Missouri	265	23
Rhode Island	265	23	Hawaii	265	23
Maryland	264	25	Oklahoma	265	23
Texas	264	25	Delaware	264	26
Delaware	262	27	Arizona	262	27
Kentucky	261	28	Florida	261	28
California	260	29	California	260	29
South Carolina	260	29	Kentucky	260	29
Georgia	259	31	West Virginia	259	31
New Mexico	259	31	Tennessee	259	31
Florida	259	31	New Mexico	259	31
Tennessee	258	34	North Carolina	258	34
West Virginia	258	34	Arkansas	258	34
North Carolina	258	34	Texas	258	34
Hawaii	257	37	Alabama	258	34
Arkansas	255	38	Georgia	256	38
Alabama	251	39	South Carolina	256	38
Louisiana	249	40	Louisiana	254	40
Mississippi	246	41	Mississippi	249	41
Dist of Colum	234	42	Dist of Colum	230	42

SOURCE: National Center for Education Statistics 1993. U. S. Bureau of the Census 1993. Calculations by ERS.

Observations About Rank Order Data

There are a number of interesting observations about the rank order data contained in Figure 6. The actual NAEP-92 state average scores ranged from a low of 234 to a high of 283 — a mere 49 point spread (out of a test total of 500 points) among the 42 states that participated in the 1992 NAEP state trial tests. Twenty-nine of the 42 states had NAEP average scores that were the same as one or more other states, and as many as 4 states had identical scores. Moreover, 20 states had only a 10 score point spread from 257 to 267 out of a maximum of 500 points. Such compact clustering of test scores makes state rankings virtually meaningless.

To illustrate the potential harm that could come from states being ranked on such compact data, especially over time, notice in Figure 6 that Ohio, Michigan, Oklahoma, and Virginia all have an average math score of 267 and therefore are tied for the rank of 18 among the 42 states. A shift of a *single* point would have dropped any one of the four states to rank 21; this shift, or even greater shifts, could have *easily* happened merely by chance or by non-school-related causes. For example, in Ohio, the standard error of the 267 mean (which is 1.5 score points, as shown in the Appendix on page 23) indicates that there is about a 50/50 chance that Ohio could have ranked 21 rather than 18 — or about the chance of a toss of a coin.

Variations in State Participation

Because participation in the NAEP state assessments is voluntary on the part of individual states, the number of states participating in each assessment could change from one year's assessment to the next. This in-and-out possibility could cause substantial fluctuations in a state's rankings that were unrelated to any changes in the state's educational programs or practices.

For example, take the four states listed above that ranked 18th, each with an average of 267, on the NAEP-92 math assessment. Suppose that all state average scores remained the same on the next NAEP state assessment, but three states that had scores above 267 in NAEP-92 decided not to participate in the next assessment. All four states would move from rank 18 among 42 states to rank 15 among 39 states. If, however, three states that did not participate in NAEP-92 decided to enter the next assessment and had scores higher than 267, the four states would drop from rank 18 among 42 states to rank 21 among 45 states.

Major Effects of Non-Response Bias

Another factor that can greatly affect state average scores on the NAEP, and therefore affect state rankings, was addressed by the Panel on the Evaluation of the NAEP Trial State Assessment Project appointed by the National Academy of Education. In its 1993 report, the NAE Panel analyzed the problem of differences in the *initial participation rates* of different states — that is, the percentage of schools from the initial sample in the state that agree to participate in the NAEP Trial State Assessment. The NAE Panel found that:

> In 1990, there were only two states with initial school participation rates below 85 percent; in the 1992 TSA, one-third of the states were in this category. The lowest 1992 participation rates were for Maine, which recruited only 62 percent of its originally sampled schools in grade 8 and 58 percent in grade 4. ...
>
> The potential seriousness of the low initial participation rates in some states is underscored by the Panel's finding that higher performance on the assessment was associated with lower initial participation rates. This finding was replicated in both the fourth- and eighth-grade samples, suggesting that the result was not due to chance. Furthermore, the Panel's analyses suggest that the finding was not due to one or two aberrant states. *Thus, there is a concern that states with low initial participation rates might have inflated results on NAEP, and the Panel finds some of the states' initial participation rates to be too low for accurate reporting of their 1992 TSA results* [emphasis in original] (National Academy of Education 1993, 100).

Such major non-response problems place in doubt the accuracy of current *rankings* and *comparisons* of the relative quality or proficiency of state educational programs based on NAEP-92 Trial State Assessment math score averages. In addition, the possibility of major fluctuations in initial school participation from one NAEP assessment to the next raises important concerns about the *reliability* of NAEP state test score averages as accurate and meaningful measures of *changes* in state *rankings* and *comparisons* over time.

Other Supporting Analyses

The findings of this analysis are supported by similar findings of other analysts. William W. Cooley, director of Pennsylvania Educational Policy Studies at the University of Pittsburgh, in an analysis of only three demographic variables — percent of residents not high school graduates, percent of students in poverty, and percent of single-parent homes — on NAEP-92 Trial State math scores, found that:

> Over 75 percent of the state variations in the math means can be explained by the nature of the populations being served by the schools in those states. Therefore, one clearly cannot use NAEP math score comparisons to make accurate inferences about the relative quality of the math programs in these 42 states (Cooley 1993, 29).

Richard M. Wolf, of the Department of Measurement, Evaluation, and Statistics, Teachers College, Columbia University, in analyzing the predictive effects of economic and demographic variables on the NAEP-90 Trial State math test scores, concluded:

> The evidence presented here clearly indicates that such differences cannot be rejected. If three readily available indicators, two that reflect general characteristics of a state's population and one that reflects a measure of a state's wealth, can predict average state NAEP performance so well, then what policy relevance can be obtained from state-by-state comparisons on the NAEP tests? (Wolf 1992, 12).

Relationship of Demographic Variables to District-by-District Comparisons

William W. Cooley has also examined the relationship of demographic variables to differences in test score averages among *school districts*. Cooley analyzed the average scores of the 500 Pennsylvania school districts on the state's mandated Test of Essential Learning and Literacy Skills (TELLS). He found that three demographic variables — percent of residents not high school graduates, percent of students in poverty, and percent of single-parent homes — yielded a negative multiple correlation of .78 with the school districts' TELLS scores. He concluded:

> This means that over 60 percent of the variation in the average student performance among these school districts can be explained by those three simple census factors, leaving only about 40 percent to be explained by all other possible factors, including other demographic variables besides these three.
>
> In other words, comparing districts on such a statewide test reveals more about the *difficulty of their educational task* than about the quality of their educational program [emphasis added] (Cooley 1993, 28).

Such findings regarding the strong effects of demographic variables on the variation in student achievement test score averages among *school districts* within a state are important because some policy makers are urging the expanded use of NAEP data to compare and rank *local school districts* on their average NAEP test scores.

State NAEP Score Adjustment Controversy

To provide "fairness" in state NAEP rankings, it has been proposed that the state NAEP test scores be adjusted statistically "to reflect differences among the states in school resources and in the ethnic, economic, and other characteristics of their student populations" (Viadero 1994, 1).

One method of adjustment would call for analyzing how a state might fare on the assessment if its population mirrored that of the nation as a whole. Another method would be to look at a state's scores as if the nation's population had the same demographic characteristics as the state. Other factors that might be used to adjust state NAEP scores include those reflecting

student "opportunity to learn" factors, such as state differences in per pupil expenditures and other measures of school resources.

Emerson J. Elliott, Commissioner of Education Statistics, U.S. Department of Education, says:

> This is an important issue, and it can't be washed away by saying the only thing the statistical agency should do is report results for the overall population (Viadero 1994, 18).

However, critics of the idea of statistical adjustments to NAEP state scores charge that such changes to state scores "would implicitly concede that poor children cannot be expected to do as well in school as their more affluent peers" (Viadero 1994, 1).

Edward Roeber, director of state assessment programs for the Council of Chief State School Officers, makes this observation about the idea of adjusting state NAEP test scores:

> Rather than saying we have standards we want all students to meet, these kinds of efforts literally have given us the idea that poor children cannot learn. The implied message is, "you have lots of poor children, so your scores should be lower" (Viadero 1994, 18).

Chester E. Finn, Jr., a member of the National Assessment Advisory Board, declares that proposals to statistically adjust state NAEP test scores are:

> ... probably the worst idea I've encountered in 10 years of closely watching NAEP. Once you start fiddling with the numbers you can ... show anything you like, and then you begin to lose public confidence (Viadero 1994, 18).

An example of the complicated and controversial problems that can occur when NAEP Trial State Assessment data are statistically adjusted to account for state demographic variables occurred recently in New Jersey. When the state teachers union, the New Jersey Education Association (NJEA), learned that the state ranked 14th in mathematics test scores among the 42 states that participated in the NAEP-92 math assessment (see Figure 6, page 10), NJEA officials commissioned a study in hopes that "the results would show that the state's public school teachers are doing a good job" (*Education Week* 1994, 4).

NJEA commissioned Howard Wainer, a researcher for the Educational Testing Service, to do an analysis *independent* of the ETS federal NAEP assessment program. The analysis found that when factors such as race and the number of limited-English-proficient students were factored in, New Jersey moved from rank 14 to rank 4 among the 42 participating states (*Education Week* 1994, 4).

It is *technically feasible* to make mathematical adjustments in the NAEP Trial State Assessment scores that *statistically* take into account the effects of various factors related to the states' NAEP score rankings. There is evidence, however, that such adjustments to state NAEP test scores would be highly controversial and perhaps create more problems than the adjustments would solve.

Measuring the Difficulty of the Educational Task

In view of the finding that 89 percent of the state differences in NAEP-92 Trial State Assessment mathematics test score averages can be explained by variations in four demographic variables over which schools have no control, it is important to ask: What *are* the differences in the NAEP-92 Trial State math scores actually measuring?

Rather than measuring differences in the quality or proficiency of the states' educational programs, the NAEP-92 Trial State average math scores appear to more accurately reflect differences in what William Cooley has termed the *difficulty of the educational tasks* confronting the various states.

State Data Used in This Analysis

The state-by-state data for each of the demographic variables used in calculating the predicted state scores and state rankings used in this analysis are shown in Figure 7 on page 16.

Summary and Conclusions

This analysis has addressed the important question of whether state average test scores from the NAEP-92 Trial State Assessment in mathematics should be used for the purpose of *comparing* and *ranking* the quality of mathematics instruction among the various states. Other uses and purposes of NAEP state or national assessments are not addressed in the analysis.

To accept a specific factor as a *valid* measure of the relative quality or proficiency of mathematics instruction among the states, one must be able to reject with reasonable confidence other possible factors influencing average test score differences, such as economic or demographic variables.

This analysis found that 89 percent of the variation in state average test scores on the NAEP-92 Trial State Assessment in mathematics can be explained by the combined effects of four demographic variables — number of parents living at home, parent(s)' education, community type, and state poverty rates.

Figure 7.—State Demographic Category/Level Data Used in this Analysis for 1992 NAEP Mathematics Assessrnent Participants

	Parent(s)' Education		Parent(s) Living at Home		Commu. Type	Poverty
State	Not High School Graduate	High School Graduate Only	One Parent Living at Home	Neither Parent Living at Home	Student Living in Disadvantaged Urban Commu.	State Poverty Rate, 1992 (Ages 5-17)
Average for 42 States and DC	8%	25%	21%	3%	9%	19.6%
Alabama	13%	29%	24%	3%	16%	23.5%
Arizona	10%	21%	22%	3%	14%	21.9%
Arkansas	11%	31%	21%	4%	5%	20.8%
California	10%	17%	22%	4%	19%	22.1%
Colorado	6%	21%	21%	2%	10%	10.9%
Connecticut	6%	22%	19%	2%	17%	16.5%
Delaware	6%	30%	24%	3%	0%	11.2%
Dist. of Colum	9%	29%	47%	8%	67%	31.8%
Florida	8%	24%	25%	3%	17%	21.6%
Georgia	11%	30%	25%	3%	10%	27.8%
Hawaii	6%	25%	21%	4%	16%	15.9%
Idaho	7%	19%	15%	2%	5%	20.2%
Indiana	8%	32%	20%	2%	11%	13.2%
Iowa	4%	25%	16%	2%	3%	14.7%
Kentucky	15%	32%	20%	3%	12%	23.1%
Louisiana	10%	30%	25%	4%	19%	32.4%
Maine	6%	26%	17%	2%	2%	16.5%
Maryland	6%	25%	23%	3%	13%	16.0%
Massachusetts	7%	21%	21%	2%	23%	18.2%
Michigan	6%	26%	23%	3%	19%	17.9%
Minnesota	3%	22%	14%	1%	0%	17.0%
Mississippi	13%	29%	27%	4%	6%	30.6%
Missouri	8%	29%	21%	3%	12%	18.6%
Nebraska	4%	24%	17%	2%	6%	14.2%
New Hamp	6%	24%	17%	2%	0%	9.3%
New Jersey	7%	23%	19%	3%	24%	13.0%
New Mexico	11%	26%	22%	3%	6%	27.7%
New York	6%	23%	23%	2%	16%	23.4%
North Carolina	10%	27%	24%	3%	5%	23.7%
North Dakota	3%	19%	13%	1%	0%	12.9%
Ohio	7%	32%	23%	2%	17%	18.3%
Oklahoma	8%	26%	20%	3%	5%	19.4%
Pennsylvania	7%	30%	19%	2%	15%	14.6%
Rhode Island	8%	22%	20%	2%	12%	19.6%
South Carolina	9%	31%	23%	4%	6%	27.7%
Tennessee	12%	29%	24%	3%	7%	17.1%
Texas	16%	21%	22%	3%	18%	23.2%
Utah	3%	15%	14%	1%	5%	11.2%
Virginia	9%	24%	21%	3%	13%	14.3%
West Virginia	13%	33%	19%	3%	10%	31.7%
Wisconsin	5%	28%	19%	1%	5%	13.8%
Wyoming	5%	23%	17%	2%	10%	12.3%

SOURCE: National Center for Education Statistics 1993. U. S. Bureau of the Census 1993. Calculations by ERS.

Since test score averages among the states are so strongly affected by four demographic factors over which schools have no control, NAEP-92 Trial State Assessment math scores are shown *not* to be valid measures to use for the purpose of *comparing* and *ranking* states according to the relative quality or proficiency of the states' educational programs.

In view of the finding that 89 percent of the state test score differences can be explained by four demographic variables over which schools have no control, it is important to ask: What *are* the differences in the NAEP-92 Trial State math scores actually measuring? Rather than measuring differences in the quality or proficiency of the states' educational programs, the NAEP state average scores appear to more accurately reflect differences in the *difficulty of the educational tasks* confronting the various states.

Analysts looking at other assessment data for local school districts within states have also found a strong relationship between demographic variables and student average scores on state tests.

This analysis also found that, in addition to the demographic influences on NAEP-92 Trial State Assessment math scores, certain *non-demographic* factors — including major variations in initial school participation rates within states — may substantially influence NAEP state test score averages and thus affect state *rankings* and state *comparisons*. The possible strong influence of such *non-demographic* variables adds an important reason for *not* considering or using state NAEP scores as *reliable* indicators of the relative quality or proficiency of state educational programs.

Proposals have been made to statistically adjust state NAEP scores to reflect more fairly the effects of state variations in factors such as school resources, ethnic groups, and specific demographic characteristics of the student population. However, debate on this idea has shown that such statistical adjustments would be complicated and highly controversial.

The findings of this analysis regarding the major impact of demographic factors on state NAEP assessment scores indicate the need to focus attention, efforts, and resources toward helping states address the *difficulty of the educational tasks* they face, rather than using NAEP state assessment data to inappropriately and unfairly *compare* and *rank* the states on presumed differences in the *quality* of their educational programs. □

Technical Note

Data on the percentage of students in each state falling within four categories — living in a disadvantaged urban community, the state poverty rate for children ages 5-17, parent(s) who did not graduate from high school or did not continue their education beyond high school, and only one parent or neither parent at home — were placed into a multiple regression equation with 1992 state NAEP mathematics scores as the dependent "y" variable to be predicted.

The resulting equation was used to generate the predicted scores based on the state demographic variables. The equation generated is:

y = 303.223+-.758*Par. Not High School Grad.

+-.01*Par. High Sch. Grad. Only +-.928*One Par. at Home

+-2.926*Neither Par. at Home +.152*Dis. Urb. Com.

+-.28*State Pov. Rate, Ages 5-17

Appreciation is expressed to ERS Research Specialist Gene Nguyen for his assistance in the preparation of this analysis.

References

Cooley, W. W. 1993. "The Difficulty of the Educational Task: Implications for Comparing Student Achievement in States, School Districts, and Schools." *ERS Spectrum* Vol. 11, No. 3 (Summer 1993): 27-31.

Drazen, S. 1992. *Student Achievement and Family and Community Poverty: Twenty Years of Education Reform.* Paper presented at the Annual Meeting of the Eastern Psychological Association, Boston, Massachusetts, April 2-5, 1992. (ED 346 234).

Education Week. April 27, 1994. "N.J. Study Shows Race, Language Have Impact on NAEP Scores." P. 4.

Haertel, E. H. April 1991. "Should the National Assessment of Educational Progress Be Used to Compare the States?" *Educational Researcher* Vol. 20, No. 3: 17.

Koretz, D. M. April 1991. "State Comparisons Using NAEP: Large Costs, Disappointing Benefits." *Educational Researcher* Vol. 20, No. 3: 19-21.

Lapointe, A. E., Mead, N. A., and Askew, J. M. 1992. *Learning Mathematics.* Princeton, NJ: Educational Testing Service.

Lazer, S. 1992. *Learning About the World.* Princeton, NJ: Educational Testing Service.

Mullis, I. V. S., J. A. Dossey, E. H. Owen, and Gary W. Phillips. 1993. *NAEP 1992 Mathematics Report Card for the Nation and the States: Data from the National and Trial State Assessments.* Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

National Academy of Education. 1993. *The Trial State Assessment: Prospects and Realities. The Third Report of The National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: 1992 Trial State Assessment.* Stanford, CA: Author.

National Center for Education Statistics. 1993. *Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States.* Washington, DC: U.S. Government Printing Office.

National Center for Education Statistics. 1990. *National Education Longitudinal Study of 1988: A Profile of the American Eighth Grader.* Washington, DC: U.S. Government Printing Office.

Pallas, A. M., G. Natriello, and E. L. McDill. 1989. "The Changing Nature of the Disadvantaged Population: Current Dimensions and Future Trends," *Educational Researcher* Vol. 18, No. 5: 16-22.

Phillips, G. W. April 1991. "Benefits of State-by-State Comparisons." *Educational Researcher* Vol. 20, No. 3: 17-19.

Robinson, G., and D. Brandon. 1992. *Perceptions About American Education: Are The Based on Facts?* Arlington, VA: Educational Research Service.

U.S. Bureau of the Census. 1993. *Table 25. Poverty Status by State and Ten Large Metropolitan Areas in 1992.* Washington, DC: Author.

Viadero, D. 1994. "Proposal to Adjust NAEP Scores for Diversity Muddled." *Education Week* (March 9, 1994): 1, 18.

Wolf, R. M. 1992. "What Can We Learn from State NAEP?" *Educational Measurement: Issues and Practice* Vol. 11, No. 4: 12.

BEST COPY AVAILABLE

Appendix

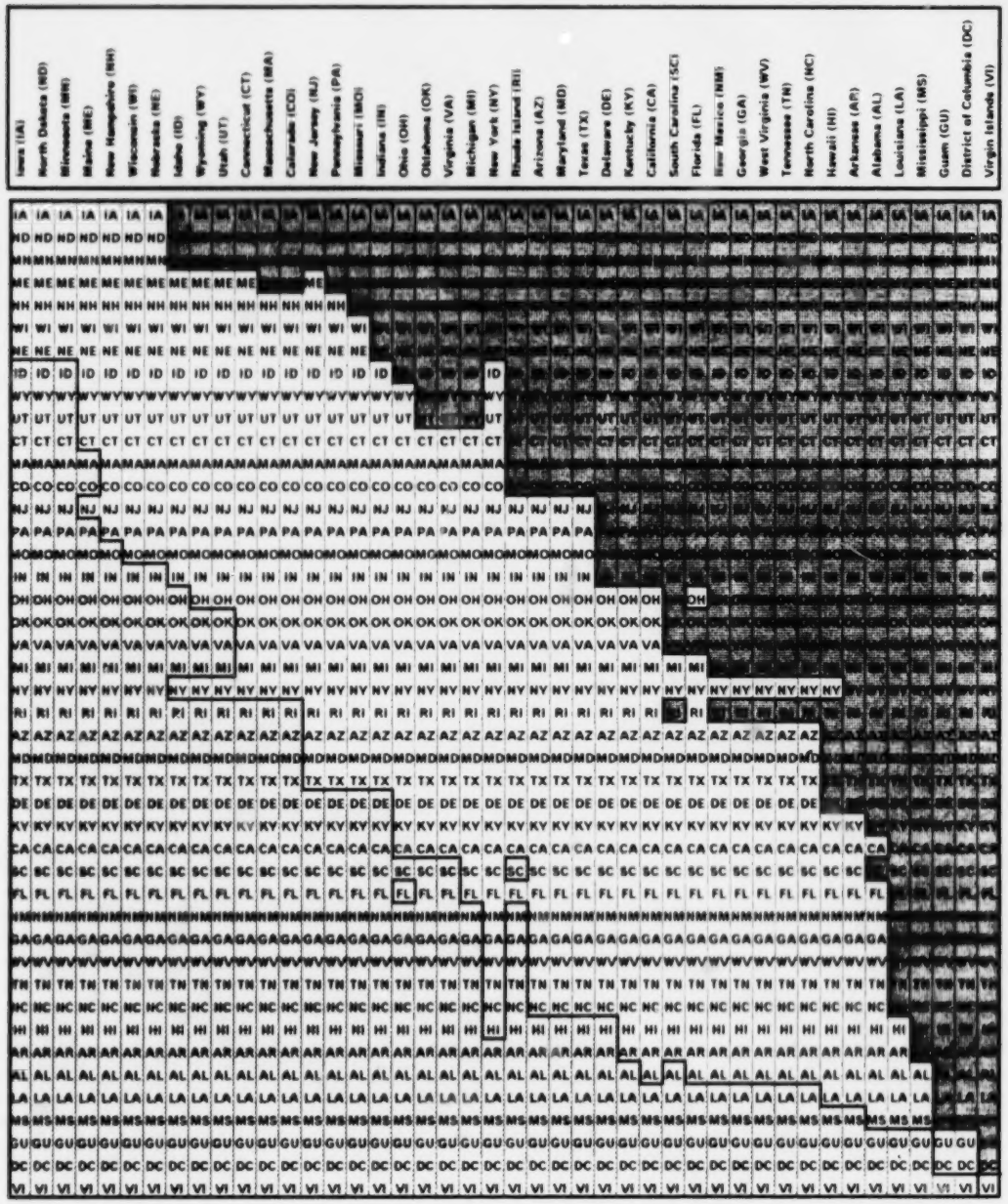
The figures on the following two pages are reproduced from the report of the 1992 NAEP Trial State Assessment in mathematics, titled *NAEP 1992 Mathematics Report Card for the Nation and the States: Data from the National and Trial State Assessments*. This report was conducted under contract by the Educational Testing Service and published by the U.S. Department of Education's Office of Educational Research and Improvement.

These two figures are shown here as examples of the ways in which the NAEP-92 Trial State Mathematics Assessment report encourages and helps readers to make state comparisons and to obtain rankings that the report identifies as indicating "significant differences between states in average mathematics proficiency."

FIGURE 3 **Comparisons of Overall Mathematics Average Proficiency 1992 Grade 8**

THE NATION'S REPORT CARD

INSTRUCTIONS: Read ***down*** the column directly under a state name listed in the heading at the top of the chart. Match the shading intensity surrounding a state postal abbreviation to the key below to determine whether the average mathematics performance of this state is higher than, the same as, or lower than the state in the column heading.



■ State has statistically significantly higher average proficiency than the state listed at the top of the chart.

□ No statistically significant difference from the state listed at the top of the chart.

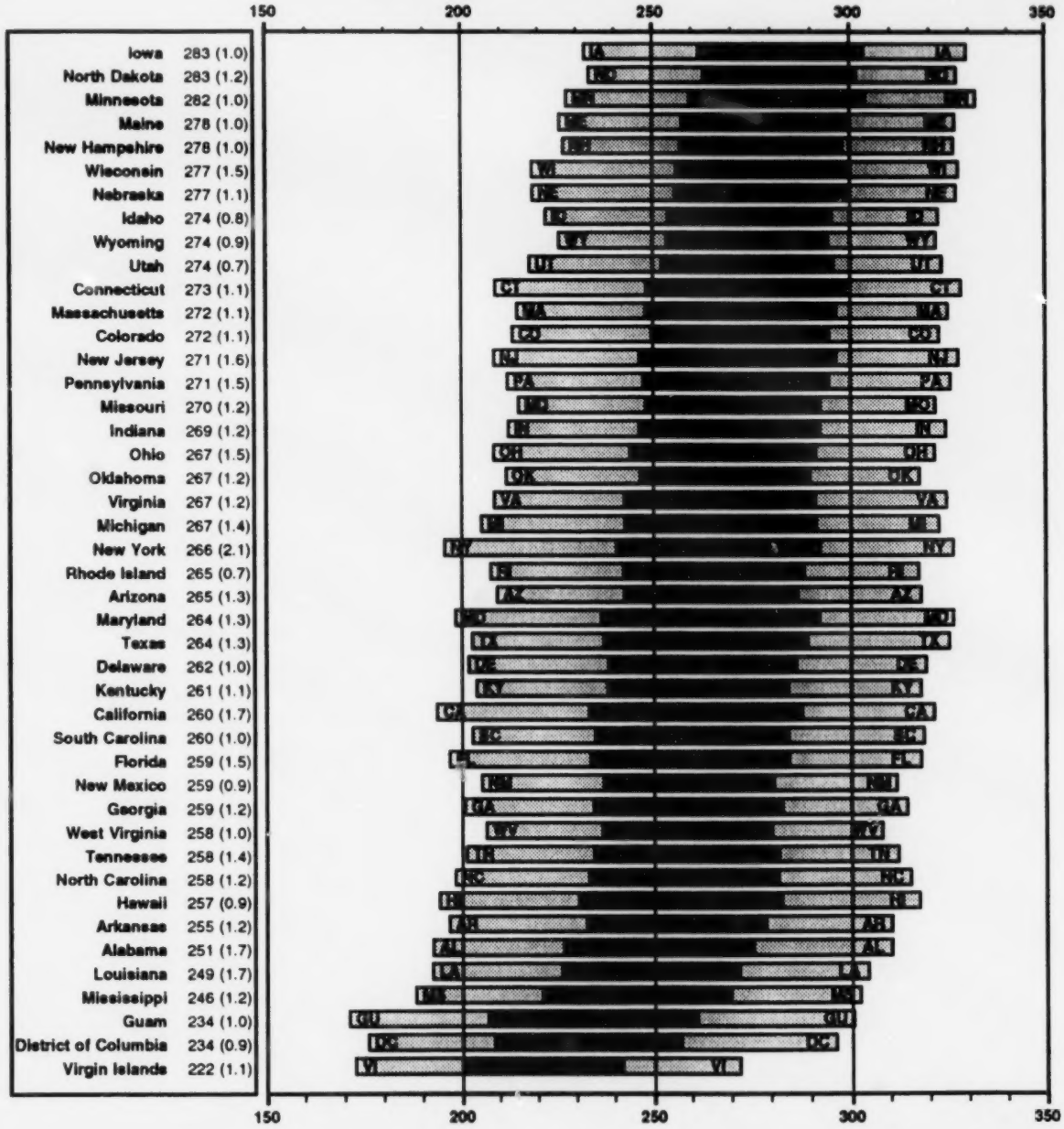
□ State has statistically significantly lower average proficiency than the state listed at the top of the chart.

The between state comparisons take into account sampling and measurement error and that each state is being compared with every other state. Significance is determined by an application of the Bonferroni procedure based on 946 comparisons by comparing the difference between the two means with four times the square root of the sum of the squared standard errors.

BEST COPY AVAILABLE

FIGURE 4 **Distribution of Overall Mathematics Proficiency Organized by Average Proficiency 1992 Grade 8**

THE NATION'S REPORT CARD



The center ***darkest*** box indicates a simultaneous confidence interval around the average mathematics proficiency for the state based on the Bonferroni procedure for multiple comparisons. Center boxes that do not overlap indicate significant differnces between between states in average mathematics proficiency. The ***darker shaded*** boxes indicate the ranges between the 25th and 75th percentiles of the mathematics proficiency distribution, and the ***lighter shaded*** boxes the ranges between the 5th to 25th percentiles and the 75th to 95th percentiles of the distribution.

Educational Research Service

The Information Source for School Decisions

Educational Research Service is an independent, nonprofit agency established to serve the research and information needs of the nation's local school districts, their administrative staffs, teachers, school boards, and the public. ERS is sponsored by the following seven national associations of school administrators:

- American Association of School Administrators
- American Association of School Personnel Administrators
- Association of School Business Officials
- Council of Chief State School Officers
- National Association of Elementary School Principals
- National Association of Secondary School Principals
- National School Public Relations Association

ERS services are available to local school districts, state and local associations of school administrators and school boards, regional educational service agencies, state departments of education, university departments of educational administration, and other related organizations. Subscription rates are graduated according to school district size or type of agency, and are available on request.

For one annual fee, ERS subscribers receive a broad range of resources, including: access to the ERS On-Call Information Service for all administrators and board members in the district; a copy of each new ERS study or report (12 to 15 each year); the monthly *ERS Bulletin*, summarizing research produced by other agencies; the quarterly *ERS Spectrum*, the journal of school research and information; access to the ERS School Operations Information Bank, including the thrice-yearly *School Operations Information Exchange;* and ERS Local School Budget Profiles Service.

For more information about ERS services, write, call, or fax:
Dr. Glen Robinson, President, or Dr. John Forsyth, Vice President, Educational Research Service, 2000 Clarendon Boulevard, Arlington, Virginia 22201. Phone: (703) 243-2100. Fax: (703) 243-8316.